Searching for Effective Teachers with Imperfect Information

Douglas O. Staiger
Dartmouth College

Jonah E. Rockoff
Columbia Business School

May 2010*

* We thank Thomas Kane for substantial comments and suggestions. Financial support was provided by Institute of Education Sciences, US Department of Education grant #R305C090023.

I. Overview

The No Child Left Behind (NCLB) Act has focused increased attention on recruiting and retaining effective teachers. To improve the quality of the teacher labor force, NCLB encouraged better screening at the time of recruitment, requiring that states hire only "highly qualified" teachers with certain minimum qualifications (having a bachelor's degree, completing an approved certification program, passage of a test of basic content knowledge). However, the sorts of teacher qualifications that are observed at the time of recruitment are, at best, only weakly related to the impact that a teacher has on student achievement. More recently, there has been growing interest in retaining teachers who have proven to be effective in the classroom, linking tenure decisions to their students' achievement gains. However, while there is considerable heterogeneity in teacher effectiveness, student achievement gains are an imprecise measure of teacher performance that could result in arbitrary tenure decisions.

Over the past four decades, empirical researchers—many of them economists—have accumulated an impressive amount of evidence on teachers: the heterogeneity in teacher productivity, the rise in productivity associated with teaching credentials and on-the-job experience, rates of turnover, the costs of recruitment, the relationship between supply and quality, the effect of class size and the monetary value of academic achievement gains over a student's lifetime. In fact, now that the data needed to estimate individual teacher performance based on student achievement gains have become more widely available, teaching may be the most-scrutinized occupation in the economy. However, there have been few efforts to take stock of the whole, to draw from the now voluminous literature on teacher performance and examine the optimal design of policies for the recruitment and evaluation of teachers.

In this paper, rather than present another estimate of a particular aspect of teacher performance, we ask what the *existing* evidence implies for how school leaders would recruit and evaluate teachers. In particular, we explore how best to use imperfect measures of teaching effectiveness at hire and during the first few years of a teacher's career to recruit and retain effective teachers. We begin by outlining a simple search model, in which schools search for teachers based on noisy signals of teacher effectiveness. We then use estimates from our work with the two largest school districts in the nation (Los Angeles Unified and New York City) to calibrate the key parameters and use the model to explore a range of questions. How much can be gained from using the current imperfect measures of teacher effectiveness based on student gains? How much could be gained by extending the time before tenure in order to gather more data on teacher effectiveness? How much emphasis should school districts place on pre-service certification as opposed to performance evaluation on the job? And, finally, what is the potential value of gathering better information about teacher effectiveness, either at the time of hire or during the first few years of teaching?

The answers are surprising, if only because they are so strikingly different from current practice. For instance, most school districts grant tenure status to teachers as a matter of course after two to three years on the job. Performance evaluation is typically a perfunctory exercise and very few teachers are, at least officially, considered ineffective (see, for example, Weisberg et al. (2009)). However, given the substantial observed heterogeneity of teacher effects and the modest rise in productivity with on-the-job experience, our simulations suggest that tenure protections should be limited to those who meet a very high bar. Even with the imprecise estimates of teacher effectiveness currently available, our simulations suggest that the optimal strategy would be to sample extensively from the pool of potential teachers, and offer tenure only to a small percentage. Such a strategy could yield annual gains in student achievement similar to those seen in recent evaluations of charter schools and class-size reductions. Moreover, if we could gather better information about teacher effectiveness at the time of hire or during the first few years of teaching, then the potential gains from such a strategy would be two to three times larger.

II. A simple model of searching for an effective teacher

In this section, we outline a simple model of using imperfect estimates of teacher effectiveness to screen out ineffective teachers and maximize student achievement. Our model is analogous to standard models of job search in which there is learning about productivity on the job (Jovanovic, 1979) and builds on a model developed in more detail in Kane and Staiger (2005). Instead of a worker searching for the most productive job, we have a principal searching for the most productive teacher. In the language of search models, we assume teachers are an experience good (Mortensen, 1986) – principals can learn only so much at the time of hire, and must learn more about teacher productivity by observing performance on the job. Thus, the principal draws teachers from the applicant pool, observes noisy signals over time about teacher productivity, and decides whether to dismiss unproductive teachers and start the process over again.

The General Model

More specifically, suppose that teacher effectiveness is composed of a fixed component that varies across teachers plus a return to experience that is the same for all teachers. All teachers improve over the first few years of teaching, but some teachers are persistently more effective than others in their cohort. Teacher effectiveness is not observed directly, and the principal must search for effective teachers based on noisy signals. We assume that the process by which a principal searches for an effective teacher is fairly simple. First, the principal collects applications and gathers information about each applicant. Based on this information (the “pre-hire” signal), she chooses the most promising candidates to fill the available vacancies. Each year she gathers additional information on each new hire’s performance in the classroom (the “on-the-job” signal). The principal may dismiss a teacher at the end of each year until the teacher reaches tenure (usually after the 3rd year). If she chooses to dismiss a teacher, or if a teacher chooses to leave for other exogenous reasons, she must start the process over again. Teacher turnover is costly because of the time and effort involved in dismissing and recruiting a new teacher, and because replacement teachers will have no prior experience. Ultimately, the principal tries to manage this process in a way that maximizes average teacher effectiveness in her school.

The optimal strategy for this type of search model has the reservation property: at the end of each year, the principal dismisses a teacher if their expected effectiveness given the information to date lies below a reservation value. The reservation value rises with time on the job, because the option value of waiting to dismiss a teacher declines as the principal accumulates better information over time. In other words, to avoid unnecessary turnover the principal may choose to wait a year before dismissing a teacher who she believes is "below the bar" so long as there is a reasonable chance that her beliefs could change. Thus, the principal dismisses teachers whose expected effectiveness lies below a bar that increases with teacher experience. Overall, the principal must set the bar to trade off the short-term cost of replacing an experienced teacher with a rookie against the long-term benefit of selecting only the most effective teachers.

An Illustrative Example

A simplified version of this general model yields an analytical solution that illustrates the tradeoffs facing a principal. Suppose that a teacher's classroom performance each year (Y_t) is the sum of a persistent teacher effect (μ), an error term that is independent across years (ε_t), and a negative "rookie" effect ($\beta<0$) in the first year ($t=1$), where both μ and ε_t are normally distributed with mean zero. Further assume that there is no pre-hire signal, so that each new hire is a random draw from the teacher distribution. Finally, suppose that new hires must be either dismissed or tenured at the end of their first year. This simplifies the solution by eliminating any option value of waiting. If not dismissed, there is some known probability that a teacher will leave voluntarily each year.

In this simple version of the model, the principal will grant tenure if a teacher's classroom performance in the first year exceeds a minimum cut-off ($Y_1>c$). The cut-off (c) for tenure is chosen to maximize the average productivity of the entire workforce. The workforce consists of two groups of workers: rookies in their first year of teaching, whose expected performance is just β, and teachers who survived the tenure cut-off, whose expected performance is $E(\mu \mid Y_1 > c)$. Therefore, the productivity of the workforce ($\bar{Y}$) is equal to:

$$(1) \qquad \bar{Y} = \pi\beta + (1-\pi)E(\mu \mid Y_1 > c)$$

Where the proportion of rookies in the workforce (π) is an increasing function of both the exogenous turnover rate and the tenure cutoff below which rookies are dismissed.[1] Thus, raising the cutoff increases the expected productivity of teachers reaching tenure, but at the cost of raising the proportion of the workforce who are rookies.

Maximizing $\bar{Y}$ with respect to c yields the following simple first-order condition determining the choice of the optimal value of c:

$$(2) \qquad E(\mu \mid Y_1 = c) = \bar{Y}$$

[1] In steady state, $\pi = \dfrac{1}{1+\sum_{t=2}^{T}\delta_t \Pr(Y_1 > c)}$, where δ_t represents the exogenous proportion of teachers who would be voluntarily (if given tenure) still teaching in year t.

The above expression has a fairly straightforward interpretation. The expression on the left is the productivity of the *marginal* teacher, whose performance was at the cut-off c. The expression on the right is the productivity of the *average* teacher (including both tenured teachers and rookies). The principal sets the cut-off, c, where the productivity of the *marginal* teacher is equal to the productivity of the *average* teacher. In other words, this decision rule tells principals to keep only the rookies who are expected to be better than the average teacher.

Imagine if this were not true. That is, suppose the marginal teacher were less productive than the average teacher. The district could raise performance by raising its standard by a small amount. Likewise, if the marginal teacher were more productive than the average teacher, then the district could raise average performance by lowering the cut-off and adding one more above-average teacher. This result is analogous to the usual result that average costs are minimized at the point where marginal cost equals average cost.

The above first order condition, in combination with the definition of average productivity in Equation 1, has a number of implications for the determinants of the cut-off level of performance required for tenure. First, a more negative rookie effect (β) lowers the average productivity of the workforce, which in turn lowers the optimal cutoff for tenure. Put simply, the value of experience raises the cost of dismissing experienced teachers. Similarly, a high exogenous turnover rate raises the fraction of rookies in the workforce (π) and lowers the average productivity of the workforce, which again lowers the optimal cutoff for tenure. There is less benefit to giving tenure to highly effective teachers if they do not stay long. Finally, low variance in the teacher effect (μ) lowers the benefit of selection, and high variance in the error with which productivity is measured (ε_t) makes it more difficult to select highly effective teachers, both of which lower the optimal cutoff for tenure. There is little reason to be selective if the performance data (Y_1) cannot reliably identify important productivity differences between teachers.

In a more general model, which allows for a pre-hire signal and for more than one year of observation before tenure, there is no simple closed form solution for calculating the reservation value. However, the optimal reservation value depends on a similar set of underlying parameters: the variation in performance across teachers, the strength of the pre-hire and on-the-job signals, the return to experience, the number of years before tenure, the exogenous turnover rate, and the size of the applicant pool and magnitude of other hiring and firing costs. Therefore, we will use evidence on these key underlying parameters from New York and Los Angeles to calibrate the model, and then solve the model numerically using monte carlo methods.

III. Relevant Evidence on Teacher Effectiveness

In this section we present evidence on the key parameters needed to calibrate the search model. While we cite evidence from other work, we rely mostly on our own estimates from Los Angeles and New York City for calibrating the model. We begin with the evidence on heterogeneity in teacher productivity and the error with which it is measured. We then turn to the evidence on hiring and firing costs. Finally, we review the evidence on information available at the time of hire.

The Heterogeneity in Teacher Productivity

More than three decades ago, Hanushek (1971) and Murnane (1975) were the first economists to report large differences in student achievement in different teachers' classrooms, even after controlling for students' prior achievement and characteristics. That literature has accelerated in recent years. Especially following the No Child Left Behind Act of 2001, many states and school districts began collecting annual data on students and matching it to teachers.[2] Research has produced remarkably consistent estimates of the heterogeneity in teacher impacts in different sites. For example, using data from Texas, Rivkin, Hanushek and Kain (2005) find that a standard deviation in teacher quality is associated with 0.11 student-level standard deviations in math and 0.095 standard deviations in reading. Using data from two school districts in New Jersey, Rockoff (2004) reports that one standard deviation in teacher effects is associated with a 0.1 student-level standard deviation in achievement. Using data on high school students in Chicago, Aaronson, Barrow and Sander (2003) report that a standard deviation in teacher quality is associated with a difference in math performance of 0.09 to 0.16 student-level standard deviations.[3]

As several recent papers remind us, the statistical assumptions required for the identification of causal teacher effects with observational data are extraordinarily strong-- and rarely tested (Andrabi, Das, Khwaja and Zajonc (2008), McCaffrey et. al. (2004) , Raudenbush (2004), Rothstein (2008), Rubin, Stuart and Zannutto (2004), Todd and Wolpin (2003)). Teachers may be assigned classrooms of students that differ in unmeasured ways—such as consisting of more motivated students, or students with stronger unmeasured prior achievement or more engaged parents—that result in varying student achievement gains.

Despite these concerns, several pieces of evidence suggest that the magnitude of variation in teacher effects is driven by real differences in teacher quality. First, while the assumptions implicit in the empirical specifications used to estimate teacher effects may not always be correct, estimates tend to be highly correlated across a wide variety of specifications (Harris and Sass, 2006). Second, while most studies of teacher effects rely on assumptions regarding matching of students with teachers at the classroom level, Rivkin et al. (2005) use a completely different approach that does not rely on this assumption and find similar estimates to the rest of the literature. Finally, two studies base their estimates on teacher-student links that were

[2] The data requirements for measuring heterogeneity in teaching effectiveness are high. First, one needs longitudinal data on achievement for individual students matched to specific teachers. Second, achievement data are needed on an annual basis, to be able to track gains for each student over a single school year. (Prior to NCLB many states tested at longer intervals, such as 4th and 8th grade.) Third, panel data on teachers are required as well, to be able to track performance of individual teachers over time. Teacher-level panel data are needed to account for school-level or classroom level shocks to student achievement that contribute to the measurement error in classroom-level measures. In earlier work (Kane and Staiger (2002)), we showed that conventional estimates of sampling error can not account for the lack of persistence in school-level value-added estimates. There appear to be other school-level and classroom-level sources of error.

[3] Aaronson, Barrow and Sander report the variance in teacher quality to be .02 to .06 grade-level equivalents (adjusted for sampling error). In Table 1, they report the standard deviation in grade-level equivalents of 8th grade students to be 1.55. ($\sqrt{.02}/1.55=.09, \sqrt{.06}/1.55=.16$) Their study adjusted for sampling variation, but not for other classroom level sources of error.

randomly assigned. Nye, Konstantopoulos and Hedges (2004) re-examine data from the Tennessee STAR classroom size experiment, in which teachers were randomly assigned to classes of a given size. The differences in classroom-level student achievement that emerge within given size groups are larger than would have been expected to occur due to chance and are strikingly similar in magnitude to those estimated in non-experimental studies. Kane and Staiger (2008) study a recent experiment in LAUSD, and examine the degree to which non-experimental value-added estimates from a pre-experimental period are able to predict student achievement differences following random assignment. Students assigned to teachers with higher "value-added" during the pre-experimental period outperformed students assigned to low "value-added" teachers and, moreover, a one-point difference in pre-experimental value-added was associated with a one-point difference in student achievement following random assignment. Thus, they could not reject the hypothesis that the non-experimental estimates for individual teachers were unbiased.

Estimation Error

The estimation error in teacher impact estimates derives from at least two sources. The first is sampling variation. The typical elementary classroom may have 20 to 25 students per year (although middle and high school teachers have somewhat larger classes and typically teach multiple sections). With samples of such modest size, naturally occurring variation in the make-up of a teacher's classroom from year to year will produce variation in a teacher's estimated impact. However, volatility in teacher (and school) impacts exceeds that predicted by sampling error alone (Kane and Staiger (2002), Kane, Rockoff and Staiger, 2008) The source of this second type of error (actually, non-persistent variation in teacher impact estimates) could include a broad range of factors influencing the measured achievement gains of groups of students—such as a locally virulent flu-season, interactions between a specific teacher's lesson plans and the test used in a given year, a dog-barking in the parking lot on the day of the test or more mysterious forces in the broad category of "classroom chemistry."

For our purposes, any non-persistent variation in a teacher's measured impact on student achievement represents estimation error. One simple approach to estimating the proportion of variance due to non-persistent sources is to study the correlation in estimated impacts across classrooms taught by the same teacher. If a teacher's estimated impact, Y_t, represents the sum of a persistent component, μ, and an uncorrelated non-persistent error, ε_t, then the correlation between Y_t and Y_{t-1}, represents an estimate of the reliability of the teacher-level estimate in any given year. Table 1 reports the standard deviation in estimated teacher effects, the estimated reliability (i.e. correlation across classrooms) and implied standard deviation in true teacher impacts ($\sigma_{\mu,}$) for teachers in two school districts, Los Angeles Unified and New York City. When reported in terms of the student-level standard deviation in test scores in a given grade and subject, the standard deviation in estimated value-added for teachers was remarkably similar in the two districts, with estimates in both math and English language arts in the narrow range from .23 to .27. Although the estimated reliability of teacher impacts was higher in math than in English language arts, and higher in Los Angeles than in New York City, all suggest that there is considerable error (i.e. volatility) in the teacher impact estimates. Indeed, more than half of the variation in estimated impacts in math and English Language Arts are non-persistent. The standard deviation of the persistent teacher effect is between .12 and .19, similar to that found in the previous literature discussed above.

Learning on the Job and Cost of Teacher Turnover

Table 1 also reports the degree to which average teacher impacts on student achievement differ from experienced teachers during the first few years on the job in these same two districts. In both Los Angeles and New York, teacher impacts on student achievement appear to rise rapidly during the first several years on the job and then flatten out. This is a finding which has now been replicated in a number of states and districts (Rivkin, et al. (2005), Clotfelter et al. (2006), Harris and Sass (2006), Jacob (2007)), The lion's share of the increase in average teacher impact occurs during the first year of a teacher's career. When assigned to a first-year teacher, the average student gains .06 to .08 standard deviations of achievement less than observably similar students assigned to experienced teachers. However, the achievement gains of students assigned to second-year teachers lagged those in more experienced teachers' classrooms by only .01 to .04 standard deviations. In Los Angeles, students of third-year teachers saw gains comparable to those of more experienced teachers, while there was a small difference for third year teachers in New York (.01 to .02 standard deviations).

Therefore, every time a school district loses an experienced teacher with two or more years of experience and is forced to hire a novice teacher, the students assigned to the novice teacher lose roughly .10 standard deviations in student achievement. To attach an approximate dollar value to that cost, one needs to estimate of the value of academic achievement over the course of a students' lifetime. There is a long tradition in labor economics estimating the relationship between various types of test scores and the earnings of early-career workers.[4] For instance, Murnane, Willett and Levy (1995) estimate that a one-standard deviation difference in math test performance is associated with an 8 percent hourly wage increase for men and 12.6 percent increase in for women. These estimates may understate the value of test performance, since the authors also control for years of schooling completed. Neal and Johnson (1996), who do not condition on educational attainment, estimate that an improvement of one standard deviation in test performance is associated with 18.7 and 25.6 percent increases in hourly wages for men and women, respectively. Kane and Staiger (2002) estimated that the value of a 1 standard deviation gain in math scores would have been worth $110,000 at age 18 using the Murnane, Willett and Levy estimates and $256,000 using the Neal and Johnson results. In sum, the economic cost of lost academic achievement when replacing an experienced elementary teacher with a novice would be roughly .10 standard deviations times $110,000 to $256,000 value per standard deviation times 20 students per class—or $220,000 to $512,000.

This cost of lost academic achievement dwarfs any other costs of teacher turnover. Milanowski and Odden (2007) carefully studied costs of teacher recruitment and hiring in a large urban Midwestern school district. They estimate total costs of roughly $8200: recruiting costs per vacancy of $1100 in central office staff time and $2600 in school-level staff time, plus $4500

[4] Of course, the cross-sectional relationship between tested achievement and earnings may could overstate the causal value of academic achievement. However, while there have been attempts to estimate the causal value of schooling, we are not aware of estimates of the causal value of academic achievement.

for the cost of training a new teacher. In addition, it is worth noting that some of these costs will be defrayed by the lower salaries typically earned by new teachers.

Evidence on the Ability to Select Effective Teachers during the Hiring Process

Can school leaders discern between effective and ineffective teachers at the point of recruitment? This is an important question, since better selection at the front end reduces the need to be selective among the set of teachers a principal hires. Unfortunately, there is scant evidence that principals can effectively separate effective and ineffective teachers when they make hiring decisions. Indeed, this notion is supported by the simple fact that most of the variation in teacher effects occurs within schools.

In our view, one of the most interesting pieces of evidence on this topic comes from a natural experiment which occurred in California in the late 1990s. Beginning in the academic year 1996-1997, the state of California provided cash incentives to school districts to keep class sizes in kindergarten through third grade to a maximum of 20 children. In order to take advantage of the state incentive, school districts throughout the state dramatically increased hiring of new elementary teachers. Figure 1 reports the hire dates of elementary school teachers working for LAUSD in May of 2003. As is dramatically apparent, there was a large increase in the number of elementary school teachers hired between 1996 and 1997. In the years before 1997, the district hired 1200 to 1400 elementary school teachers per year, but in 1997 LAUSD nearly *tripled* the number of elementary school teachers it hired, to 3,335.[5]

If the district were able to effectively discern teacher effectiveness in the hiring process, we would have expected a large increase in hiring to have had a negative impact on the average effectiveness of the teachers hired. This effect would have been heightened by the fact that nearly every other school district in California was on a hiring spree because of the same state law. However, despite the size of the hiring bubble, there is little evidence that the average teacher hired in 1997 was any worse than those hired in the years immediately before 1997. Figure 1 plots the coefficients on dummies for teacher hiring cohort, in a regression specification examining student achievement of those teaching in grades 2 through 5 in Los Angeles during 2001 through 2004. As is apparent in Figure 1, there is little evidence that the average effectiveness of the 1997 hiring cohort was any different from the much smaller cohorts hired in prior years.[6] By 2001, roughly two-thirds of both the 1996 and 1997 hiring cohorts were still employed by the district (thus, there is little evidence to suggest that there was any differential selective attrition for the larger cohort). Although the specification used to estimate teacher impacts included controls for baseline scores and other student characteristics (gender, race/ethnicity, federal lunch program participation, English language learner status), there is virtually no difference in the types of students to which the cohorts had been assigned.

[5] We coded someone as being hired in the 1997 academic year, if they were hired between July 1, 1996 and June 30, 1997. We defined the other academic years in the same way.

[6] This evidence runs counter to the prevailing wisdom among policy analysts, that it was a decline in the average quality of the teaching force that accounts for the failure to see an increase in achievement in California resulting from the class size reduction. (Bohrnstedt and Stecher, 2002)

Other evidence on this issue comes from decades of work in which researchers have tried, unsuccessfully, to link teacher characteristics (observable to both researchers and principals) to student outcomes (see reviews by Hanushek (1986, 1997) and Jacob (2007). With the exception of teaching experience, there is little to suggest that the credentials commonly used to determine teacher certification and pay are related to teachers' impacts on student outcomes. There are some studies which find that a teacher's academic background (e.g., college GPA, SAT test scores) is related to student outcomes, but Ballou (1996) finds that teaching applicants with strong academic records are no more likely to be hired by school principals.

More recent work suggests that selecting teaching candidates who are likely to be effective is difficult, but not impossible. For example, several studies have estimated the impact of novice teachers recruited under the Teach for America (TFA) program (Decker et al. (2004), Boyd et al. (2006), Kane, Rockoff and Staiger (2008)). TFA is extraordinarily selective, drawing applicants from the top universities in the country and offering positions to only a small fraction of the thousands of individuals who apply. Decker et al. (2004)use random assignment to estimate the impact the TFA program in elementary schools and find that students assigned to TFA members scored 2 percentile points (0.095 standard deviations) higher in math and no higher in reading than those assigned to other teachers. Using non-experimental data from New York City, Kane et al. (2008) find positive impacts of TFA teachers in math of .02 standard deviations and no statistically significant effect in English language arts. Boyd et al. (2006) report comparable results, also using data from New York City.

More evidence comes from studies collecting data on recently-hired novice math teachers in New York City. Rockoff, Jacob, Kane and Staiger (2008) collected information on a number of non-traditional predictors of effectiveness including teaching specific content knowledge, cognitive ability, personality traits, feelings of self-efficacy, and scores on a commercially available teacher selection instrument and then used these to predict a teacher's impact on math achievement. When the variables were combined into the two primary factors summarizing cognitive and non-cognitive teacher skills, those teachers who were one standard deviation higher on either the cognitive or non-cognitive factor seemed to raise student achievement in math by only .033 student-level standard deviations. (Those who were 1 standard deviation higher on both measures were estimated to raise achievement by .066 standard deviations.) Rockoff and Speroni (2010) examine the achievement of students assigned to teachers recruited through an alternative certification program—the New York City Teaching Fellows—and ask whether achievement gains were higher for students assigned to teachers rated as more attractive candidates by the certification program's interview protocol. They find no significant relationship with English language arts test scores and a small positive relationship with math test scores: a one standard deviation in interview score was associated with .013 standard deviations higher math achievement gain.

Summary of the Evidence

The evidence just reviewed can be succinctly summarized in four points. First, the standard deviation across teachers in their impact on student achievement gains is on the order of 0.1 to 0.2 student level standard deviations. Second, value added estimates of teacher effectiveness have reliability of 30% to 50%. Third, the primary cost associated with dismissing an experienced teacher is that the average student gains will be .06 to .08 standard deviations of achievement less with a first-year teacher. Fourth, it is difficult to reliably identify effective teachers at the time of hire.

IV. Implications for How We Should (and Should Not) Search for Effective Teachers.

In this section, we use the estimates cited above to calibrate our model and simulate average teacher productivity under various scenarios. We begin with a benchmark analysis of the optimal search strategy given the current quality of information available and assuming that tenure is given after the first year. We then compare this benchmark case to a number of policy relevant alternatives. How much better would schools do if we extended the time until tenure? Alternatively, what if we required schools to accumulate at least 2 or 3 years of performance data before dismissing a teacher? Finally, what would be the benefit of collecting better information that would let us more accurately identify effective teachers (either at hire or in the first years of teaching)? This final question is motivated by a variety of efforts under way to develop better methods to identify effective teachers.

Benchmark Simulations

The benchmark simulations use the evidence from the previous section to set the key parameters of the model. We assume that districts do not observe any useful pre-hire signal, and that the on-the-job signal is an annual value added measure. We set the standard deviation of the persistent teacher effect (in student-level standard deviation units) equal to 0.15, and the reliability of the value added measure (the ratio of the persistent variance to total variance) equal to 40%. For the return to experience, we assume that a first and second year teacher's value added is -0.07 and -0.02 student standard deviations below the value added of teachers in their 3rd year or higher. All of these values lie in the middle of the estimates reported for LA and NYC in Table 1. We ignore the direct costs of hiring a new teacher, since the evidence in the prior section suggested that these are small relative to the cost associated with rookie teachers' lower value added. Finally, we assume a maximum teaching career of 30 years and an exogenous turnover rate of 5 percent, which is approximately the proportion of experienced teachers who leave the LA and NYC districts each year.

We begin with the simple case in which the principal must either dismiss or tenure a teacher after their first year of teaching based on just one year of value added data. Figure 2 reports the expected steady-state impact of dismissing a given proportion of teachers (the bottom axis) on value added of the average teacher (left axis, solid line) and on the proportion of the teacher workforce who are in their 1st year of teaching (right axis, dashed line).

The implications of Figure 2 are stark. First, the simulation suggests that there are substantial gains from using value added information to dismiss ineffective teachers. For example, if the principal dismissed the bottom third of first-year teachers with the lowest value added rather than dismissing no teachers, then the average value added among teachers in the school would increase by a bit over 0.04 in the long run. Second, the simulation suggests that the principal should set a very high bar for tenure, and dismiss about 80% of teachers after their first year. This aggressive strategy would raise the average value added of teachers in the school to just over 0.08. Third, the simulation suggests that the optimal dismissal rate of 80% would result in roughly one quarter of the workforce being novice teachers. Currently, only about 10% of the teaching force in LA and NYC is made up of novice teachers. This implies that the districts would have to more than double the hiring of new teachers to accommodate this aggressive strategy.

While these results are surprising relative to current practice, there are a number of clear reasons why it is optimal for principals to dismiss a large proportion of novice teachers. The main reason is that differences in teacher effects are large and persistent, relative to the short-lived costs of hiring a new teacher. As a result, even unreliable performance measures such as value added can identify substantial and lasting differences across teachers. Since the typical teacher getting tenure will teach for ten years or more, the benefit from setting a high tenure bar will be large. Of course, such unreliable measures make mistakes. But the long-run cost of retaining an ineffective teacher far outweighs the short-run cost of dismissing an effective teacher. Moreover, because of the uncertainty at the time of hire, new teachers have considerable option value; for every five new hires, one will be identified as a highly effective teacher and provide many years of valuable service.

There are, nevertheless, a number of potential reasons that our simulations may overstate the benefits or understate the costs of such an aggressive tenure policy. First, we may have understated the hiring and firing costs facing a principal. However, even if we double the difference in value added between rookies and experienced teachers (which corresponds to an additional hiring cost of well over $100,000 in terms of foregone future student earnings), the optimal dismissal rate remains over 75%. Second, we may have understated turnover rates among tenured teachers, especially if principals focus on their own school (rather than the district as a whole) and highly effective teachers are more likely to move to other schools. Similarly, principals may discount the future more highly because of their own likelihood of leaving the school, or because they believe that teacher effects will not persist into the future (although the evidence suggests otherwise). However, if we double the annual turnover rate from 5% to 10%, the optimal dismissal rate remains over 70%. Third, we may have understated the cost of recruiting teachers, since new hires would presumably demand higher wages to compensate for the substantial risk of being dismissed. This is particularly true if we continue to require costly up-front teaching-specific training. However, even a doubling of current teacher salaries would not be enough to offset the benefits of an aggressive dismissal policy, since a .08 annual increase in student achievement is worth more than $100,000 per teacher. Finally, we may have overstated the reliability of value added measures, because even the persistent component of value added is an imprecise measure of what principals really value in a teacher. However, as we will discuss below, even if we cut the reliability of value added in half, from 40% to 20%, the optimal dismissal rate remains over 70%.

The Impact of Changing the Time to Tenure Review

In Table 2, we use simulations to evaluate how changing the time until tenure review affects the optimal dismissal rate and the average value added of teachers. The first column repeats the results from our benchmark simulations in which dismissal could only occur at the end of the first year. The next three columns allow the principal to delay tenure review until the 2nd, 3rd, or 4th year, and to gather more information about teacher effectiveness before making her decision regarding dismissal. The next three columns *require* a delay in tenure review for 2-4 years, so that dismissal can occur only after multiple years of value added data are available to the principal.

Not surprisingly, giving a principal the option of waiting to gather more information produces some benefits. As a principal is given the option to wait until year 2, 3, or 4 to make a decision, overall dismissal rates rise by a few percentage points. The principal would still dismiss two-thirds of new hires after the first year, but she would wait to dismiss some teachers for whom there is a reasonable chance that an additional year of data could lead to a better decision. Average value added rises to about 0.10 standard deviations with the possibility of delaying tenure review to the 4th year, with most of the gain coming from delaying tenure until the 2nd year.

In contrast, *requiring* principals to delay tenure review (i.e., removing the option of dismissal until year 2, 3, or 4) would lead to lower average teacher value-added, relative to the baseline case. Essentially, this policy forces principals to retain low-performing teachers additional years, and this outweighs the benefits of the additional information the principal would obtain by waiting to see additional years of performance data. Note that this policy also leads to fewer teachers being dismissed overall, since the option value of hiring a new teacher (who may turn out to be ineffective and must be retained for several years) has fallen.

Obtaining More Reliable Measures of Teacher On-The-Job Performance

Figure 3 shows how changing the reliability of the on-the-job signal affects the optimal timing of tenure (regions delineated by dotted lines, labeled at top), the optimal dismissal rate (right axis, dashed line), and the resulting value added of the average teacher in the school (left axis, solid line). For these simulations, we assumed that the principal could only dismiss teachers at tenure time (T). Our baseline simulation corresponds to a reliability of 40% (0.4) in this figure.

Many districts are currently engaged in efforts to improve the reliability with which they can measure teacher performance, through the use of additional information from classroom observation, student work, and student or parent surveys. Figure 3 suggests that more reliable measures of teacher performance are quite valuable. Relative to the baseline simulation, a measure with perfect reliability would nearly double the gains from selecting effective teachers (to 0.14 standard deviations) while having little impact on the proportion of teachers dismissed. If districts relied on performance measures that were less reliable than our baseline case, the gains from selecting effective teachers would be reduced, and it would become optimal for the principal to wait longer before dismissing a teacher. Interestingly, the proportion of teachers dismissed does not decline much until the reliability of the performance measure drops below 5%. Even very weak signals of teacher performance eventually identify differences between teachers that swamp the cost of hiring more inexperienced teachers.

Obtaining More Reliable Information at the Time of Hire

We have assumed that principals have no useful information at the time of hire. This implies that radical increases in hiring rates (as required by a dismissal rate of 80%) do not affect the quality of new hires – each individual is a random draw from the applicant pool. But many districts and principals put substantial effort into screening and interviewing new hires, suggesting that even small amounts of information at the time of hire may be valuable.

Figure 4 shows how changing the reliability of the pre-hire signal affects the optimal dismissal rate (right axis, dashed line), and the resulting value added of the average teacher in the school (left axis, solid line). For these simulations, we assumed that the principal could only dismiss teachers after the first year (T=1). We also assumed that the pool of potential applicants was ten times the number needed to replace teachers leaving through exogenous turnover, corresponding to estimates that NYC and LA currently have about 10 applicants for each position. Our baseline simulation corresponds to a reliability of 0 in the pre-hire signal, at the far left in this figure.

Figure 4 suggests that pre-hire information on teacher effectiveness is potentially quite valuable. Compared to having no information at the time of hire, a perfect pre-hire signal with 100% reliability would nearly triple the value added of the teacher workforce, and eliminate the need to dismiss teachers after hire. This should not be surprising, since observing effectiveness perfectly at the time of hire allows one to select the best candidates out of the applicant pool and avoid discovering later that some teachers were ineffective. More interestingly, even a low reliability signal (20%) at the time of hire doubles the value added of the teacher workforce relative to the benchmark case with no pre-hire information. However, access to a pre-hire signal does not eliminate the need to dismiss additional teachers after hire – so long as there is remaining uncertainty about teacher effectiveness among the teachers that are hired there will be a benefit to dismissing additional teachers after observing classroom performance.

V. Conclusion

In the debate over how to improve teaching quality in public schools, some have made claims regarding the statistical properties of currently available measures of teacher effectiveness and their usefulness. For example, in reference to a proposed initiative to measure teacher effectiveness using student test scores, the head of the New York City teachers' union (Randi Weingarten) stated: "There is no way that any of this current data could actually, fairly, honestly or with any integrity be used to isolate the contributions of an individual teacher." She also went on to claim that "Any real educator can know within five minutes of walking into a classroom if a teacher is effective" (New York Times (2008)). We do not believe that either of these statements is well supported by existing research.

The goal of our analysis was to quantify the potential value of information on teacher effectiveness. Our analysis provides a useful (although certainly not complete) framework for thinking more systematically about how best to use the information that is available, and about the returns to collecting better information on teacher effectiveness in the future.

Most school districts attempt to screen out ineffective teachers at the point of hiring and then do little to screen out ineffective teachers afterwards. This strategy is consistent with institutions that require individuals to make large occupational-specific investments prior to becoming teachers. However, research suggests that there is currently little information available at hire that allows us to discriminate between effective and ineffective teachers. Our results suggest a very different approach to raising the effectiveness of the teaching workforce. They imply that one could identify much larger differences between teachers by observing a single year of teaching performance and retaining only the highest-performing teachers. Despite the fact that current estimates of teacher performance are fairly noisy, they can still be used aggressively to identify effective teachers and increase the overall quality of teaching in public schools. This approach is consistent with an initial process of hiring that is not selective – and in particular does not require teachers to make costly educational investments prior to being hired.

Our simulations suggest that using existing information on teacher performance to aggressively select teachers would yield substantial annual gains in academic achievement of around 0.08 student level standard deviations. These are comparable to the annual test score gains found in recent experimental evaluations of charter schools (Hoxby and Murarka (2009), Abdulkadiroglu et al. (2009)) and comparable to the estimated annual impact of reducing class size in early elementary grades found in Project STAR (Krueger (1999)). There may be other uses of this information that we did not consider in our analysis, such as for performance-based pay or targeted professional development, which would yield even larger gains. Systematically exploring the potential gains from these other uses would certainly be valuable.

Of course, there is no reason that districts should be content with the imperfect information currently available on teacher performance. Our analysis also suggests that there are substantial returns to investing in better information about teacher effectiveness, both at the time of hire and in the first few years on the job. Other measures of teacher performance, such as parent or principal evaluations, classroom observations, or even "teacher tryouts" in summer school classes, may be useful. While incorporating such measures into teacher evaluation is a promising development, the general message of our analysis would remain – that such measures should be used aggressively to identify and retain only the best teachers early in their career.

References

Aaronson, Daniel, Lisa Barrow, and William Sander (2007) “Teachers and Student Achievement in the Chicago Public Schools,” Journal of Labor Economics, 25(1): 95-135.

Abdulkadiroglu, A., Angrist, J., Dynarski, S., Kane, T.J., Pathak, P. (2009). “Accountability and Flexibility in Public Schools: Evidence from Boston’s Charters and Pilots”. NBER working paper #15549.

Andrabi, Tahir, Jishnu Das, Asim Ijaz Khwaja, and Tristan Zajonc (2009 “Do Value-Added Estimates Add Value? Accounting for Learning Dynamics,”

Ballou, Dale (1996) “Do Public Schools Hire the Best Applicants?” Quarterly Journal of Economics, February 1996

Bohrnstedt, G., and B. Stecher (2002). *What We Have Learned about Class Size Reduction in California,* Palo Alto, Calif.: California Department of Education.

Boyd, Donald Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff “How Changes in Entry Requirements Alter the Teacher Workforce and Affect Student Achievement,” Education Finance and Policy 1(2): 176-216.

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor (2006) “Teacher-Student Matching and the Assessment of Teacher Effectiveness,” NBER Working Paper No. 11936.

Decker, Paul T., Daniel P. Mayer and Steven Glazerman. (2004) “The Effects of Teach For America on Students: Findings from a National Evaluation,” Mathematica Policy Research Report No. 8792-750, June 9, 2004.

Hanushek, Eric A. (1971) “Teacher Characteristics and Gains in Student Achievement: Estimation using Micro Data,” American Economic Review Vol. 61, No. 2, pp. 280-288.

Hanushek, Eric A. (1986) “The Economics of Schooling: Production and Efficiency in Public Schools,” Journal of Economic Literature, 24(3): 1141-1177.

Hanushek, Eric A. (1997) “Assessing the Effects of School Resources on Student Performance: An Update,” Educational Evaluation and Policy Analysis,19 (2): 141-164

Harris, Douglas N., Sass, Tim R. (2006). “Value-Added Models and the Measurement of Teacher Quality”. Unpublished manuscript, April 2006.

Hoxby, C.M., Murarka, S. (2009). “Charter Schools in New York City: Who Enrolls and How They Affect Their Students’ Achievement”. NBER working paper #14852.

Jacob, Brian (2007) “The Challenges of Staffing Urban Schools with Effective Teachers,” The Future of Children 17(1): 129-154.

Jovanovic, Boyan (1979) "Job Matching and the Theory of Turnover," Journal of Political Economy, 87(5): 972-990.

Kane, Thomas J. and Douglas O. Staiger (2002) "The Promise and Pitfalls of Using Imprecise School Accountability Measures," Journal of Economic Perspectives, 16(4): 91-114

Kane, Thomas J. and Douglas O. Staiger (2005). "Using Imperfect Information to Identify Effective Teachers". Unpublished manuscript, April 2005.

Kane, Thomas J., and Douglas O. Staiger (2008) "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper 14607.

Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger (2008) "What Does Certification Tell Us about Teacher Effectiveness? Evidence from New York City," Economics of Education Review, 27(): 615–631

Krueger, Alan B. (1999) "Experimental Estimates of Education Production Functions," Quarterly Journal of Economics, 114(2): 497-532.

McCaffrey, Daniel F., J. R. Lockwood, Daniel Koretz, Thomas A. Louis, and Laura Hamilton (2004) "Models for Value-Added Modeling of Teacher Effects," Journal of Educational and Behavioral Statistics, 29(1): 67-101.

Milanowski, Anthony T., Odden, Allan R. (2007). "A New Approach to the Cost of Teacher Turnover". School Finance Redesign Project Working Paper 13.

Mortensen, D. 1986. "Job Search and Labor Market Analysis." Chapter 15 in Handbook of Labor Economics, Volume 2. Edited by O. Ashenfelter and R. Layard, North-Holland.

Murnane, Richard. (1975) The Impact of School Resources on the Learning of Inner City Children (Cambridge, MA: Ballinger).

Murnane, Richard J., John B. Willett and Frank Levy (1995) "The Growing Importance of Cognitive Skills in Wage Determination," Review of Economics and Statistics, 77(2): 251-266

Neal, Derek A. and William R. Johnson (1996) "The Role of Premarket Factors in Black-White Wage Differences," Journal of Political Economy, 104(5): 869-895

Nye, Barbara, Spyros Konstantopoulos, and Larry V. Hedges (2004) "How Large Are Teacher Effects?" Educational Evaluation and Policy Analysis,26 (3): 237-257

Raudenbush, Stephen W. (2004). "What are Value-Added Models Estimating and What Does This Imply for Statistical Practice?" Journal of Educational and Behavioral Statistics 29(1):121-129, Spring 2004.

Rivkin, Steven G., Eric A. Hanushek, and John Kain. (2005) "Teachers, Schools and Academic Achievement," Econometrica Vol. 73(2)

Rockoff, Jonah E. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," American Economic Review, May Papers and Proceedings.

Rockoff, Jonah E., Jacob, Brian, Kane, Thomas J., Staiger, Douglas O. (2008). "Can You Recognize an Effective Teacher When You Recruit One?" *Education Finance and Policy*, forthcoming.

Rockoff, Jonah E. and Cecilia Speroni (2010) "Subjective and Objective Evaluations of Teacher Effectiveness," American Economic Review, Papers and Proceedings, 100(2).

Rothstein, Jesse (2010) "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," Quarterly Journal of Economics, 125(1): 175-214

Rubin, Donald B., Elizabeth A. Stuart and Elaine L. Zanutto (2004) "A Potential Outcomes View of Value-Added Assessment in Education," Journal of Educational and Behavioral Statistics, 29(1): 103-116.

Todd, Petra E., and Kenneth I. Wolpin (2003) "On the Specification and Estimation of the Production Function for Cognitive Achievement," The Economic Journal, 113: 3-33

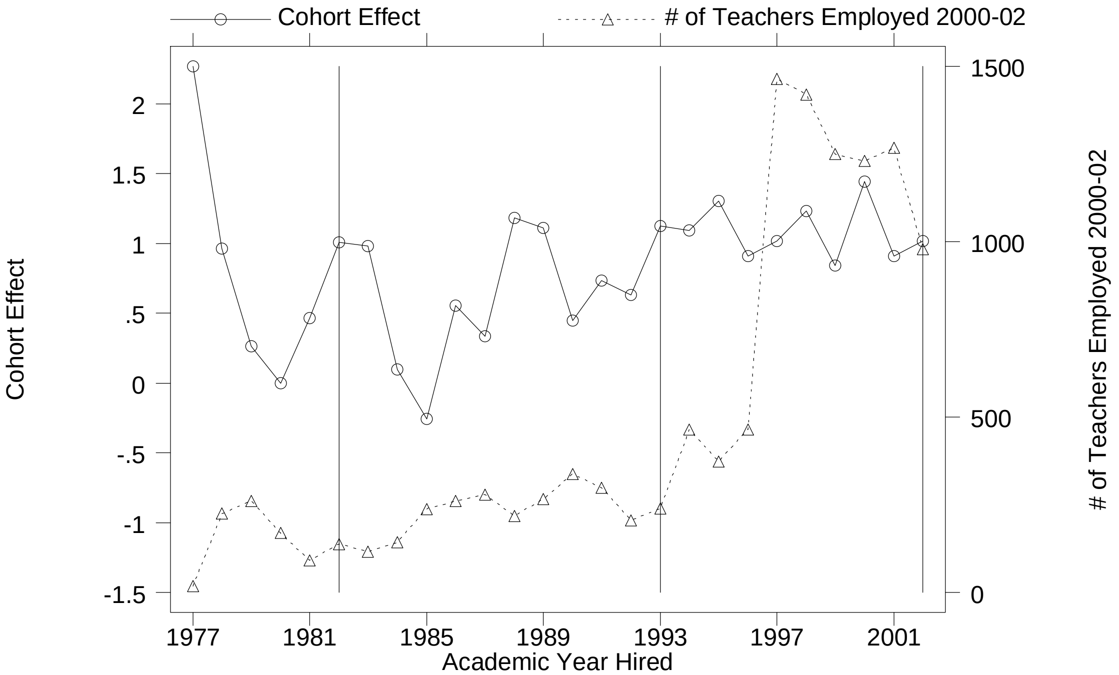
Weisberg, Daniel, Susan Sexton, Jennifer Mulhern, and David Keeling. 2009. The Widget Effect. Brooklyn, NY: The New Teacher Project.

Table 1. Evidence on Teacher Value Added From LAUSD and NYC Schools.

	Los Angeles		New York City	
	Math	ELA	Math	ELA
Variation in Teacher Value Added:				
Standard Deviation of Annual Value Added Measure	0.27	0.23	0.25	0.23
Reliability of Annual Value Added Measure	0.50	0.37	0.39	0.28
Implied Standard Deviation of Persistent Teacher Effect	0.19	0.14	0.15	0.12
Difference in value added relative to teachers with 3+ years experience				
No experience teaching (Novice)	-0.08	-0.06	-0.07	-0.07
One year of experience teaching	-0.02	-0.01	-0.03	-0.04
Two years of experience teaching	-0.01	-0.01	-0.02	-0.02

Note: Estimated using 4th and 5th graders in years 2000-2003 for Los Angeles, and 2000-2005 for New York City. ELA refers to English language arts. Reliability of the value-added measure refers to the correlation in of the value-added measure across classrooms taught by the same teacher. Teacher value-added estimated including student-level controls for baseline test scores, race/ethnicity, special ed, ELL, and free lunch status; classroom peer means of the student-level characteristics; and grade-by-year fixed effects.

Figure 1. Value added and size of the cohort in 2000-2002 plotted against the academic year in which the cohort was hired in LAUSD.



Cohort Size and Estimated Cohort Quality

Figure 2. Expected impact of dismissing a given proportion of teachers after their first year of teaching based on one year of value added data. Steady state impact on value added of average teacher (left axis, solid line) and on proportion of teacher workforce who are in their 1st year of teaching (right axis, dashed line).

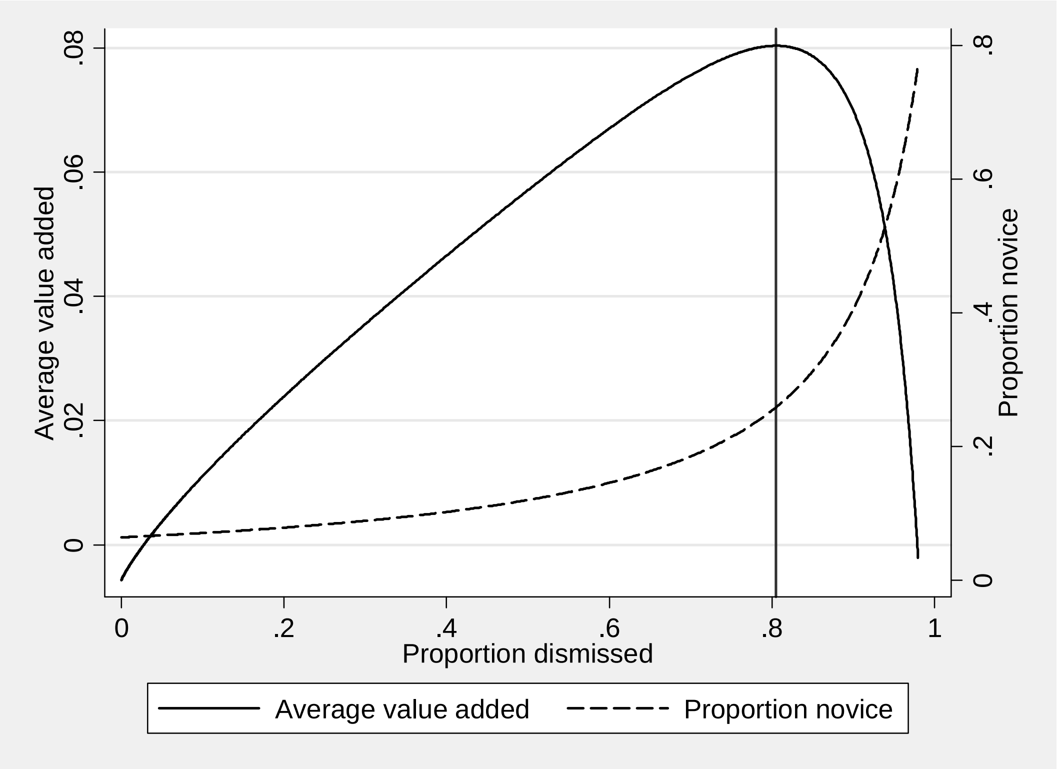


Table 2. Expected impact of delaying tenure decision on value added of the average teacher and cumulative dismissal rates.

	Baseline: Dismissal at T=1	Dismissal Allowed at Any Time Until:			Require Dismissal only Occur at Time:		
	T=1	T=2	T=3	T=4	T=2	T=3	T=4
Average Value Added	0.080	0.095	0.099	0.101	0.075	0.068	0.061
% Dismissed Overall	81%	83%	84%	84%	75%	71%	68%
% Dismissed Annually							
At T=1	81%	67%	67%	67%			
At T=2		16%	8%	8%	75%		
At T=3			9%	4%		71%	
At T=4				5%			68%

Figure 3. Impact of increasing the reliability of the annual performance measure on value added of average teacher and proportion of teachers dismissed after the optimal waiting period (T).

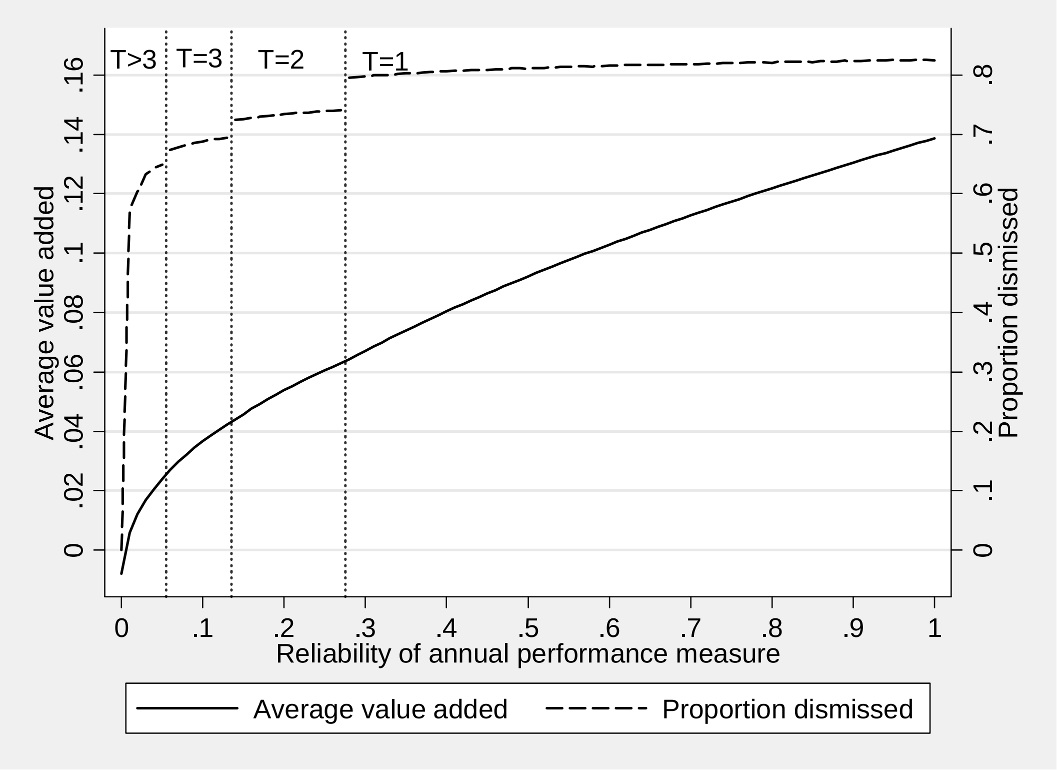
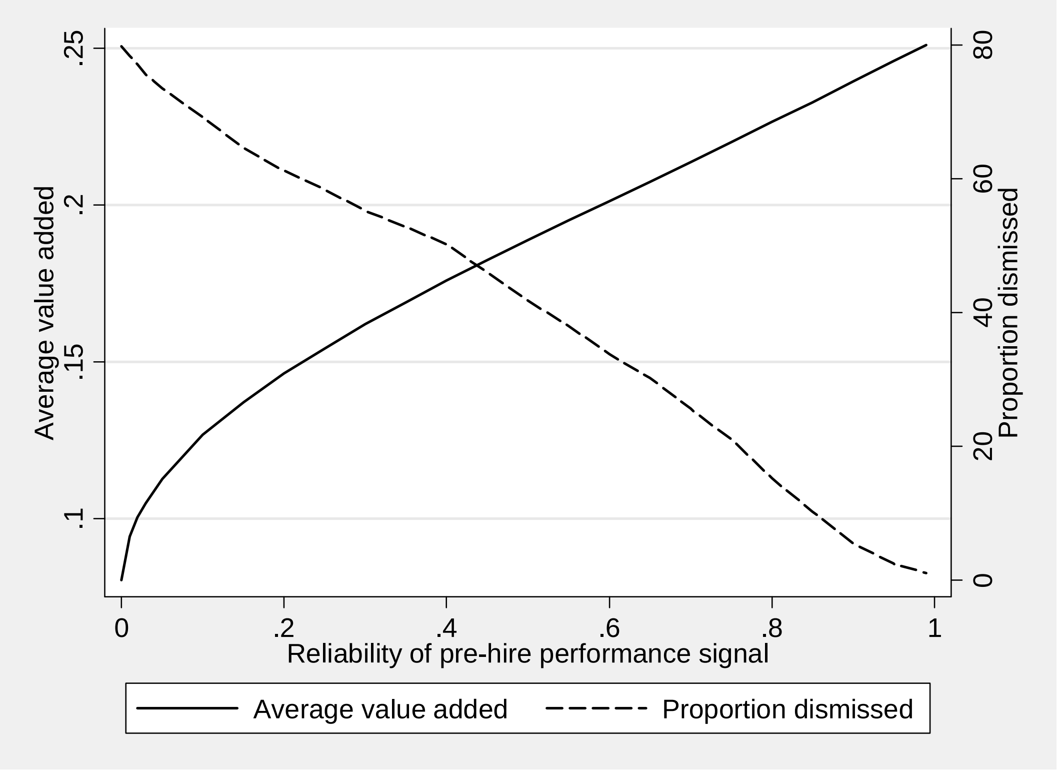


Figure 4. Impact of increasing the reliability of the pre-hire performance signal on value added of average teacher and proportion of teachers dismissed after one year.


Average value added
Proportion dismissed
Reliability of pre-hire performance signal
Average value added
Proportion dismissed